5-2009

Traffic Analysis for the Calibration of Risk Assessment Methods

Brian R. Calder
University of New Hampshire, Durham, brian.calder@unh.edu

Schwehr Kurt
University of New Hampshire, Durham

Traffic Analysis for the Calibration of Risk Assessment Methods

B. R. Calder* K. Schwehr†

Abstract

In order to provide some measure of the uncertainty inherent in the sorts of charting data that are provided to the end-user, we have previously proposed risk models that measure the magnitude of the uncertainty for a ship operating in a particular area. Calibration of these models is essential, but the complexity of the models means that we require detailed information on the sorts of ships, traffic patterns and density within the model area to make a reliable assessment. In theory, the AIS system should provide this information for a suitably instrumented area. We consider the problem of converting, filtering and analysing the raw AIS traffic to provide statistical characterizations of the traffic in a particular area, and illustrate the method with data from 2008-10-01 through 2008-11-30 around Norfolk, VA. We show that it is possible to automatically construct aggregate statistical characteristics of the port, resulting in distributions of transit location, termination and duration by vessel category, as well as type of traffic, physical dimensions, and intensity of activity. We also observe that although 60 days give us sufficient data for our immediate purposes, a large proportion of it—up to 52% by message volume—must be considered dubious due to difficulties in configuration, maintenance and operation of AIS transceivers.

1 Introduction

Assessing the risk to a vessel of transiting or anchoring in a given area is a fundamental task for the user of hydrographic data. A sufficiently nuanced analysis of risk is one way to communicate, to the user, the degree of uncertainty in the data being presented, providing a much better means to analyze and understand the completeness, accuracy and validity of the navigational product for an individual than current methods such as source or reliability diagrams, or their equivalents in electronic products (e.g., Zones of Confidence [ZOCs]).

*Center for Coastal and Ocean Mapping and NOAA-UNH Joint Hydrographic Center, University of New Hampshire, Durham NH 03824, USA. E-mail: brc@ccom.unh.edu

†Address as above; e-mail schwehr@ccom.unh.edu

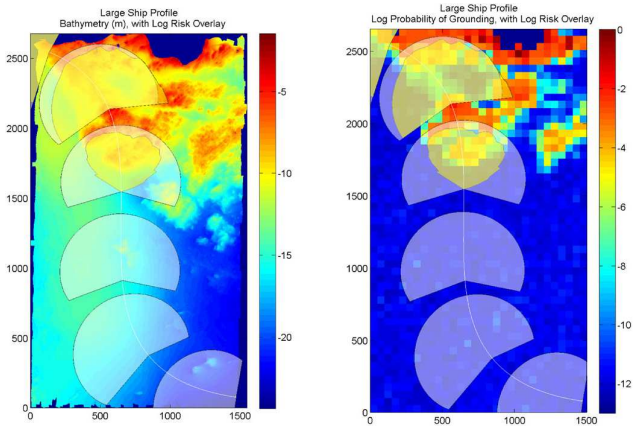

Figure 1: Example of risk assessment for a large (∼ 100 m) ship following a trajectory (in white) through shallow bathymetry (left) and associated local risk (right). The overlaid semi-transparent regions are potential motion zones within the next two minutes from each center location; the overlaid yellow figures are additional risk that would be incurred by heading in the indicated directions.

We have proposed previously methods for analysis of the risk associated with a vessel transiting through a particular area, or at anchor [1] (Figure 1). A description of risk to a vessel is useful for a number of other tasks in addition to safety of the vessel itself. We could, for example, use a risk analysis to prioritize which areas to survey (e.g., survey highest risk first)[2], determine which part of an area to survey first (e.g., prioritize survey resources to the area of highest risk), or determine when a particular survey has reached the point where further work is unwarranted (e.g., stop when the residual risk above baseline falls below a defined level).

Plausible methods for expressing the risk to a vessel in any given area, however, require more information than can be provided from the hydrographic databases typically held by Hydrographic Offices. In particular, much of the assessment of risk revolves around the behaviors of the vessel or, if the assessment is intended to describe the risk within a geographic area (such as a harbor or approach), the aggregate behavior of all of

the traffic in the area. Other issues such as preferred transit lanes, traffic control measures and local climatic conditions are also important. Unfortunately, this vital information is typically either poorly understood or difficult to obtain.

The Automatic Identification System (AIS) is a VHF ship-to-ship and ship-to-shore messaging system designed to pass vessel information. The primary goal of the initial AIS specification was to assist with safety of navigation by improving the situational awareness of all mariners. AIS is specified by International Telecommunication Union Recommendation (ITU-R) M.1371-3 [3].

The system uses two 9600 bps radio transceivers in the 160 MHz band with 1 to 12.5 W power. This RF channel limits the amount of data and gives maximum ranges from 5 to 400 km with 25–50 km being typical with most configurations. Each ship is equipped with a transceiver that operates in a Self-Organizing Time Division Multiple Access (SOTDMA) network to exchange messages. There are currently two classes of transceivers. The first, Class A, uses 12.5 W transmission power and is higher priority and more configurable than Class B, which is limited to 2 W power and must be programed by a vendor. Class B systems also transmit position reports and ship information at a lower rate. Operators of Class A equipment are expected to enter data into the device either with the Minimum Keyboard Display (MKD) or through an Electronic Charting System (ECS).

AIS has been in use since 2001, with mandatory carriage requirements for new Safety Of Life At Sea (SOLAS) class vessels since 2002-07-01. Carriage requirements are continuing to evolve with more vessel types likely to be required to carry AIS transceivers in the future. Even if it is not required, many mariners choose to add AIS transceivers to their vessels.

Operational aspects of AIS are described in International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA) Guideline No. 1028 [5]. This guideline covers AIS reception in Vessel Traffic Services (VTS), but does not detail shore side networks. Initial design requirements for shore side reception and data distribution are specified in the IALA Technical Issues document [6], IALA Guideline No. 1050 [7], and IALA Recommendation A-124 [8]. The United States Coast Guard (USCG) Research and Development Center (RDC) created an initial development network of receiving stations. This network was transitioned from development to production and moved from the RDC to the USCG Navigation Center, where it is known as National-AIS (N-AIS) Increment 1. The plans for N-AIS are illustrated in Figure 2. Increment 1 is receive-only.

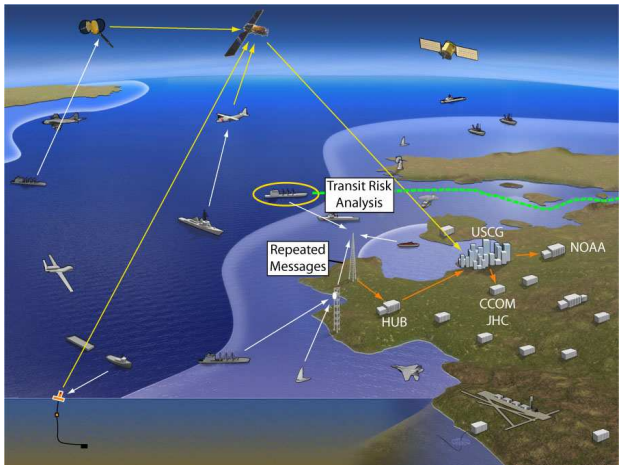

Figure 2: Schematic view of the USCG N-AIS receiving network. Vessels broadcast position and shipdata messages that are received by towers, buoys, and planes, and orbiting spacecraft. Some towers can act as repeaters, forwarding AIS messages. Messages are timestamped at a remote hub or when they reach the N-AIS command center at the USCG Navigation Center (NAVCEN) in Alexandria, VA [4]. The USCG then forwards the AIS messages through an internet based realtime distribution network. Vessels will eventually be able to send their planned route to a port authority where a risk analysis can be conducted for the transit.

Increment 2 will add the ability to send information to vessels from shore and increase the number of receivers. Increment 3 looks towards space- and buoy-based receivers to cover distant ocean areas.

AIS traffic has, in theory, the ideal characteristics to allow for calibration of risk models in many areas, with the mandatory carriage requirement for large ships making it possible to harvest data from shore in one location and use this to characterize the local traffic conditions. In addition, many smaller ships are starting to use the Class B variant of the Class A AIS transceivers seen on larger ships, and therefore may fill in a gap in coverage over time.

The difficulty in utilizing AIS data for traffic analysis is that it was, fundamentally, not designed for retrospective analysis. This means that it lacks components such as full source timestamps for all packets, or some means to reliably identify repeated packets. Since the transport is also necessarily unreliable (i.e., there is no guarantee of delivery and no attempt to support receipt tokens), and effectively stateless and memoryless, it is difficult to aggregate the data for analysis. The integrity of a great deal of the data, such as the 'Navigation Status' indicator and physical dimensions,

particularly draught, is also reliant on the user making appropriate adjustments to reflect the ship's current status. As with all manual systems, this leads to potential for confusion, misinterpretation and inappropriate configuration [9]. It is relatively simple to do rough traffic density or intensity analysis [10], but much more difficult to extract stateful synoptic descriptions of the actual behaviors of the various classes of traffic within a given area. If we wish to calibrate risk models—or even to understand the traffic in a particular area—we need to parse a little more finely.

We therefore propose a scheme for automatic processing of N-AIS traffic in a constrained area that preparses the data from the NMEA-encoded form of the AIS message format [11], applies rough spatial filtering to select a particular area, and then repackages the data in a relational database [12, 13]. The database is then pre-filtered to sanitize the contents; the intent is that after the sanitization process, the database should be able to answer queries about the traffic without the user having to take special precautions on the results (such as looking for duplicates or missing information). We next parse the database in a number of different ways to elicit the behavior of traffic in the area of interest, including the physical dimensions of ships in broad traffic classes, the duration, frequency, location and endpoints of individual transits, and patterns of arrival and departure time, vessel affinity and dock activity intensity for automatically identified areas of transit endpoint clustering. We stratify the behaviors by traffic categories and dock/anchorage areas to better tune models of particular classes, typically the large commercial traffic, which may be more important in the overall controls on traffic in the area of interest. Our goal is to establish a series of hierarchically related statistical models that can be used to simulate the behaviors of traffic in the area of interest.

We illustrate these techniques in the case of the ports of Norfolk and Hampton Roads, VA for the period between 2008-10-01 and 2008-11-30, and describe the difficulties we encountered in processing the data. We show that there are distinct patterns of duration of transits, significant class-specific clustering in the physical characteristics of ships in the area, their destinations, and their operating areas that can be exploited to simplify the resulting models. We observe that it is not possible to carry out the same analysis for all classes of traffic in the area, since many classes have either very few members or limited structure in their behavior patterns. In these cases we show that we can summarize the aggregate behavior of the class with simpler models that may not provide high fidelity modeling of the total behavior, but provide sufficient information to allow a reasonable description of the net effect.

The presence and adoption of Class B transceivers is a question of current interest. We investigate this by considering the information from San Francisco, CA during the same period, and show that although increasing numbers of Class B transceivers are observed, their relative abundance is still very small.

Finally, we provide some comments on the difficulties in processing AIS data to characterize traffic and what might be done about it in the future. We also provide some perspective on the future of traffic simulation models derived from this work, and their use in risk models to assess uncertainty for the user, prioritize areas for resurvey and calibrate survey effort in the field.

2 Methods

2.1 Pre-Processing and Database Generation

At present, there are 26 Message Identifiers denoting categories of AIS messages out of 64 total possible identifiers. Our vessel traffic analysis uses the three Class A position reports: 1) Scheduled position report, 2) Assigned scheduled position report, and 3) Special position report—response to interrogation. For this analysis, all position report types are treated as equivalent. AIS message 5, 'Static and Voyage related data' ('shipdata'), is used to obtain the vessel name, type/cargo, draught, and dimensions.

The messages were recorded from the USCG N-AIS Increment 1 network using the 'with-out duplicates' mode [4], meaning that N-AIS will only give one message even if it received at multiple towers. Messages are stored in USCG format 0—an extension to NMEA-0183 [14] that allows for additional metadata to be added to the end of each line [15]. The extension adds fields for the receive station, signal strength, slot number, time of arrival, and a UNIX UTC timestamp. Each day of messages is compressed with `bzip2` [16] to approximately 30% of the original data volume.

The next set of steps convert the position messages and shipdata reports to a relational database. We use `noaadata` [17] to convert to a SQLite [12] database in this instance.

The position messages are extracted and clipped to a bounding box spanning $76°54'\mathrm{W}, 36°12'\mathrm{N}$ to $75°6'\mathrm{W}, 37°24'\mathrm{N}$. We chose the bounding box to cover the approaches to Chesapeake Bay through to Newport News and Norfolk, VA. The retained mes-

b003669730	r003669934	r05CCPH1
b003669794	r003669935	r05RCHI1
b2003669982	r003669936	r05RMSQ1
r003381010	r003669937	r05RTUC1
r003381012	r003669938	r05SOIN1
r003381014	r003669939	r05XDCJ1
r003669931	r003669957	r3669961
r003669932	r003669959	rNDBC44014
r003669933	r01SCST1	

Table 1: N-AIS stations covering the study area. Basestations start with 'b' and receive-only stations start with 'r'.

sages are then inserted into the database by the `ais_build_sqlite.py` script in `noaadata`.

From the position messages received, a list of unique stations receiving packets from the study area is generated. The shipdata messages are then culled to those received from the 26 stations (Table 1) that provided positions within the study bounding box, and are converted to a normalized form. The raw message comes across the network in two separate NMEA strings that may be interspersed with other NMEA messages. The normal form violates the NMEA-0183 format by converting the multiple lines to one long single sentence line. The normal form is then passed to `ais_build_sqlite.py` and added to the database.

The `noaadata` software only does simple checks of the expected message sizes to check for corrupted data. However, it does not do any packet inspection to reject nonsensical content, leaving it to the data conditioning steps to dig into the packet content and look at relationships between packets to provide appropriate filtering.

2.2 Data Conditioning

In order to make the analysis of the traffic data simpler, we first pre-filter the database to condition the behavior of the data, but preserve the integrity of the data by maintaining a reserve table for each data message type, into which we insert all data points that are filtered from the primary tables. In order to maintain information on the filtering within the database, we also construct a table to hold a text description of each filter that is applied, along with a reference number. As we filter, we keep records of which unique IDs are removed from the data tables and maintain this information in a third table within the database; subroutines of the main filtering code automate this process to ensure traceability. This mechanism ensures that we can recover evidence of which data points were eliminated for each reason, and recover them if required.

As the first stage of data triage, we attempt to ensure that the data meets the requirements of the ITU standards for AIS [3]. Each ship is identified by a Maritime Mobile Service Identity (MMSI) number, a structured identifier where the first digit indicates the regional origin of the ship (or at least where it is registered), and the next two digits provide a country identified within the general region. In particular, MMSIs starting with a zero or one are not meant to be assigned to general shipping, and no ships should be using a country code not currently assigned. We therefore extract the country component of the MMSI and match it against the currently assigned code table, eliminating those position and shipdata reports that do not match the list.

Next, we apply simple sanity checks to ensure that all ship MMSIs that appear in the position messages also appear in the shipdata messages, and vice versa. This condition of bijection between the two sets can be violated if a remote tower picks up passing ship position messages but no shipdata messages or vice versa, and is not readily identifiable at the database build stage.

Third, we filter out all ships with very few position reports, since they do not contribute significantly to the traffic in a harbor. The limit of how many points are sufficient is essentially arbitrary; in this case, we cull all MMSIs that appear in fewer than 30 position messages.

Next, we consider the static data for a ship that should be consistent in all situations: the MMSI, the IMO number (if present), the ship's name and callsign. Since this data is essentially free-form, there are a number of ways in which it can be entered. The name, for example, can be arbitrarily padded with spaces, or 'end of string' characters (represented by '@'). Even when these are controlled, however, we find numerous instances of ships whose names change lexicographically, although not semantically ('MT MITCHELL' to 'MT. MITCHELL' for example), or where random bit errors in the radio transmission result in significant modifications to the data stream. While it is possible that some of these could be resolved by human interaction, it is notoriously difficult to do approximate matching of free-form text like this automatically. In order to keep the processing simple and efficient, and acknowledging that these situations are limited within the corpus, we choose to simply filter all occurrences.

Fifth, we examine the 'Ship and Cargo' specification associated with the shipdata messages, and eliminate all traffic that is not in an appropriately defined category. This filtering is approximate, since there is no

restriction on the category that the vessel sets, and indeed one vessel can belong to multiple categories at different times depending on what cargo it is carrying, and the role it is playing at the time (e.g., a tug can change category to 'towing' once it is associated with its ship, and a 'pleasure craft' could conceivably change to 'fishing' to indicate current use). Using values from the reserved sections, however, clearly indicates a misconfigured transceiver, and we filter all records from such sources.

Next, we eliminate all traffic from repeater stations. In theory, it would be possible to use repeater traffic as a means to bolster questionable traffic, or extend the range at which ships become visible but difficulties with timestamping as presented limit this in practice. It is possible that these difficulties might be resolved by detailed processing of embedded time information that occasionally appears in the data packets, but we have not pursued this in the current work.

Finally, we consider the consistency and stability of the static ship dimension data. In the shipdata messages, each ship broadcasts a length and breadth dimension (specified as two components so that the position of the primary source of position reports can be identified), as well as an estimate of the draught. While we might expect that the draught of the ship will change over time, there is little reason to believe that the overall dimensions should, at least unless there is a corresponding change in the 'Ship and Cargo' value being broadcast simultaneously. (For example, a tug that takes a barge in tow might change its dimensions to reflect the size of the combined entity.) We therefore stratify by MMSI and 'Ship and Cargo' declared, and within each group compute the mean and standard deviation of the declared length and breadth. In order to robustify the estimates, we apply a simple outlier removal algorithm by ignoring all packets that are outwith three standard deviations of the unconstrained mean, and then recomputing the mean and standard deviation. We consider the remaining packets to be dubious if there are fewer than 1% of the total packets remaining in the estimate, or if the standard deviation is greater than 10% of the mean value in either dimension, or if the length to breadth ratio is lower than 2.0 (which is extremely unusual for most traffic, but characteristic of vessels with misconfigured receivers that have swapped one or more of the measurements). Finally, we filter all ships that have draught set to zero, since they have no useful information for our current purpose (draught and derived underkeel clearance are critical for the sort of risk assessment models that are our ultimate goal). A surprising number of ships fall into this category, including many super-yachts, but also some commercial traffic where this is unexpected.

2.3 Transit Construction

The foundation for understanding the behavior of traffic in any given area is to be able to associate some semantics, which we cannot observe, with the data that we can. The most basic semantic structure is to divide the contiguous sequence of position reports into a set of transits.

The term 'transit' admits multiple possible definitions. We interpret it here to mean a sequence of position reports from a particular ship, without significant time gaps, which show some level of purposeful motion. Typically, a transit starts when the ship is first picked up by the seaward most AIS tower in the area, and finishes when the ship either goes to anchor, or ties up at a pier. (We consider the case of a ship that goes to anchor for a period and then continues to the dock to be two transits.) Along the way, it is possible that we can lose contact with a ship for a period of time, and consequently we have to allow for time gaps in the sequence but ensure that too long a time gap is considered to be a separate transit. If the ship disappears for a significant length of time, we have no guarantee that is doing the same thing when it returns as it was when it disappeared.

In theory, transit detection is trivial for AIS: the 'Navigation Status' component of the shipdata messages should provide an indication of when the ship is underway, moored, at anchor, etc. Unfortunately, however, because this information is set manually by the bridge watch, it is typically not entirely reliable. We frequently see examples where the status is changed from 'moored' to 'underway' significantly after the ship is clearly moving (and vice versa); where the status is set to 'underway' though the ship does not move more than 10 m in any direction for 60 days; or where the status changes with the bridge watch.

Similarly, it should be possible to detect consistent motion by estimating the variation of position reports over time. This requires, however, choice of an essentially arbitrary set of parameters, making it difficult to automate.

We therefore construct transits by considering the speed over ground (SOG) estimate distributed in the AIS messages, which benefits from the smoothing of the Kalman filter implemented in the GPS providing the information to the AIS transceiver. In order to satisfy the conditions laid out above, we implement the detection algorithm as a simple synchronous state machine as illustrated in Figure 3. The algorithm starts to indicate a transit when the SOG rises above station

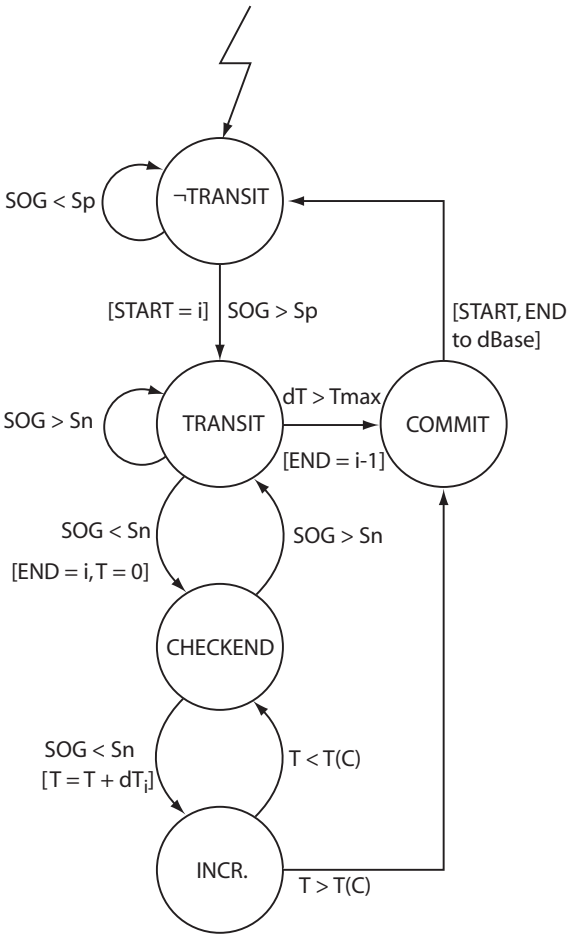

Figure 3: State diagram for the transit detection process. The state nominally advances at the rate of the position reports in the data stream, and commits detections to the current database as they are finalized in the COMMIT state.

keeping (typically $S_p = 0.5\,\text{kt}$), and indicates end of a transit when the SOG drops below a lower threshold (typically $S_n = 0.2\,\text{kt}$) for a suitable length of time (typically $T(C) = 5\,\text{min.}$). This hysteresis, and the required observation period after the SOG drops below the negative threshold, provides for sufficient robustness to allow for reliable detection of most transits in our test corpus. To provide for gap detection, we also consider a transit to be terminated by a time-gap of more than a given period (typically $T_{\max} = 10\,\text{min.}$) between subsequent samples.

We have intentionally kept the algorithm for transit detection simple so that it operates efficiently. In an attempt to provide some more semantic knowledge, however, we do carry out simple filtering on transits. In particular, we eliminate as dubious (flagged in the database) any transit that consists of a small number of position reports, or if the 2d_{rms} estimate for the transit is less than approximately three ship lengths as estimated from the shipdata reports as defined previously (a threshold derived by considering the watch-circle for a ship at anchor). As a final test, we also check the start and end points for the transit against a user-specified clip rectangle indicating the active area of interest about the analysis area. Transits without either a start or end point within the clip rectangle are marked 'clipped' in the database; those with either a start or end point inside the rectangle are marked 'trans-harbor' transits; and those with both start and end point within the rectangle are marked 'in-harbor' transits.

2.4 Transit Analysis

Having isolated individual transits per vessel, we now consider aggregation of these transits in a number of different ways that illustrate different behaviors of the traffic in a harbor as a whole.

To start the analysis we establish 12 categories of traffic according to the declared type and purpose that each ship broadcasts. The categories, given in Table 2, reflect broad classes of traffic that we might expect to have different behaviors, and therefore which we need to analyse separately in order to characterize traffic patterns in the given area.

2.4.1 Synoptic Analysis of Large Commercial Traffic

We distinguish immediately between large commercial traffic and their associated services (categories 'Cargo', 'Tanker', 'Tug', 'Towing' and 'Pilot Vessel') and the other, usually smaller, traffic. These typically comprise the majority of traffic in our areas of interest both in terms of number of vessels and number of transits, and it is therefore important to obtain a much more nuanced model of their behaviors.

We therefore select all identified transits by categories according to Table 2, and then build simple summaries of the overall behavior of the category by computing distribution estimates for the physical dimensions and the duration of the transits that the ships undertake. We use only transits that survived the clipping process in pre-filtering to eliminate problems with small transit fragments, and separately estimate behavior for in-harbor and trans-harbor transits where indicated.

Finally, we fit analytic models to the observed durations in order to parameterize them for later simula-

Ref.	Name	First Code	Last Code	Description
1	Wing-in-Ground	20	29	Experimental ships
2	High Speed Craft	40	49	Fast ferries, hovercraft and jetfoils
3	Passenger	60	69	Ships carrying paying passengers
4	Cargo	70	79	All cargo ships, including hazardous
5	Tanker	80	89	Typically oil tankers
6	Fishing	30	30	Includes fishing craft, and craft that are fishing
7	Pleasure Craft	37	37	Includes dinghies to super-yachts
8	Towing	31	32	Sometimes used by tugs while actively towing
9	Dredging	33	33	Includes dredgers on transit and active
10	Pilot Vessels	50	50	
11	Tugs	52	52	Seen both when in transit and when towing
12	Other	90	99	Vessels not in any other category

Table 2: Agglomeration categories used in the analysis. Other ship types are known, and occasionally seen in the data, but these are the principle types of significance in the test corpus. Note the potential for confusion between types of ships and their present activities, which is inherent to the design of the AIS system.

tion. It is not clear that the true distributions of the durations should be, and we therefore use a mixture of Normal, Gamma and Exponential distributions as appears appropriate to the data in the test corpus.

2.4.2 Synoptic Analysis of Other Traffic

The transits belonging to the remaining categories are typically diverse, and not nearly as consistent as that of the commercial traffic and their tenders. We therefore attempt a simpler characterization with the understanding that this provides for a less accurate portrayal of the underlying behaviors.

We compute distribution summaries of ship dimensions as above, including duration of transit, and then compute density maps for the operating range of the traffic by counting occurrence of any ships position in a regular 2D grid over the area of interest (normalized for the size of the cell). This readily identifies the primary areas of activity for the traffic, and provides a useful summary for future simulation modeling.

2.4.3 Termination Zone Detection

Ships of a particular category that all have transit endpoints with the same region have a high probability of having the same tasks and are likely to have correlated behaviors. To test this, we used the categories from Table 2 and applied the K-mean clustering algorithm [18, 19] to see where critical locations are likely located, focusing on commercial traffic (Section 2.4.1). We investigated a range of 10 to 50 clusters, with 25 clusters appearing to give a reasonable qualitative feel to vessel grouping. The best choice of clusters is likely a factor of number of ships, transits, docks, and anchorage points, and is a topic for future research. K-means has particular trouble with diffuse areas such as pilotage areas. As a result, we used the K-means clusters (or 'code book') and transit end point plots to create bounding boxes around each of the hot spots. These were plotted together on Google Earth, allowing discrimination of different operations at neighboring K-mean cluster centroids (e.g., coal versus containers), and retained as 'active zones' for further analysis.

2.4.4 Zone Activity Modeling

We take advantage of the termination zones derived from the data by category to further stratify the transits. For each zone, we select all transits which either start or stop there, and repeat the analysis as above. In addition to the distribution estimates for physical sizes and transit durations, we also compute the number of transits per day observed at the facility, the distribution of this transit density, the distribution of arrival and departure times for the ships with respect to UTC in order to elicit any preferences of time for transits at a particular zone of interest, and duration of sojourns at the dock for particular MMSIs.

3 Results

3.1 Traffic Lost to Filtering

The volume of data filtered by the various stages of the processing scheme (as outlined in section 2.2) are given in Table 3 as a percentage of the total number of packets that were observed in each message type. The biggest component here is evidently misconfigured

Filter Stage	Positions (%)	Ship Data (%)
MMSIs set to 0 or 1	0.04	N/A
MMSIs with < 30 position reports	$<$0.01	N/A
MMSIs with inconsistent static data	5.60	1.90
MMSIs for unknown country	2.71	1.62
MMSIs for unknown ship type	10.61	4.73
Repeater messages with dubious timestamps	16.41	3.92
MMSIs with dubious dimensions	16.70	7.68
Total Packets Filtered (%)	52.07	19.85
Total Packets Received	6354510	4423087

Table 3: Distribution of packets filtered by the pre-processing stage. Note the extremely high percentages lost to badly configured transceivers (unknown ship type, unstable ship dimensions).

transceivers with inconsistent static data, or claiming to be from an undefined country, or in an undefined shipping class. The total number of packets filtered for having poor dimensional stability is confusing, since this is something that might be expected to be essentially static data (with the exception of draught, which was not part of the filtering scheme). This perhaps suggests that there is either some confusion in the community on how these parameters should be set, or that there is an issue with software configuration of the transceivers where these parameters are provided automatically from some other system on the bridge.

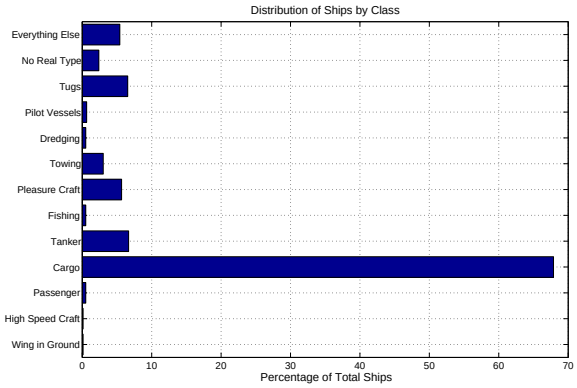

Figure 4: Distribution of ships observed in Norfolk, VA from 2008-10-01 to 2008-11-30 by shipping category from Table 2. The 'Wing in Ground' class is in fact a single (misconfigured) vessel, the BAYOU DAWN, which is in fact a tug (IMO 8955794).

3.2 Distribution of Traffic

The most basic description of traffic in the harbor is to consider the distribution of ships by category, Figure 4, or by country of registration, Figure 5. It is also illustrative to consider the distribution of detected transits by category, Figure 6. A total of 765 distinct MMSIs were observed within the analysis period, although a total of 794 combinations of MMSI and 'Ship and Cargo' values were identified, indicating that a number of vessels change status codes over time. We treat vessels with unique pairs of these parameters as separate entities because the change in code typically indicates a different behavior.

Comparison of Figure 4 and Figure 6 illustrate an immediate problem of attempting to characterize the behavior of traffic in the harbor using simpler models: the intensity of the traffic is not well defined by either metric. Thus, for example, we see many more cargo ships than tugs, but the tugs make significantly more transits in the observation period, although their transits tend to be shorter on the average (see section 3.4) and confined to particular locations within the harbor, Figure 7. In order to develop models suitable for simulation, therefore, we need to understand the correlations between the metrics so that we can assemble traffic patterns from the highest level of detail consistent with stable estimation.

3.3 Physical Dimensions of Categories

A fundamental characteristic of the traffic required for reasonable risk modeling is the physical size of the ships. Within each category, we can estimate the length, breadth and draught reported per ship, and utilize this to select 'plausible' ship dimensions for simulation. The length distributions for the categories in Table 2 are shown in Figure 8, and breadths in Figure 9. Some immediate trends are evident. First, the filtering that we have been forced to do has reduced the density of ships in the 'Wing in Ground', 'High Speed Craft', 'Passenger', 'Fishing' and 'Dredging' categories such that there are too few model ships to re-

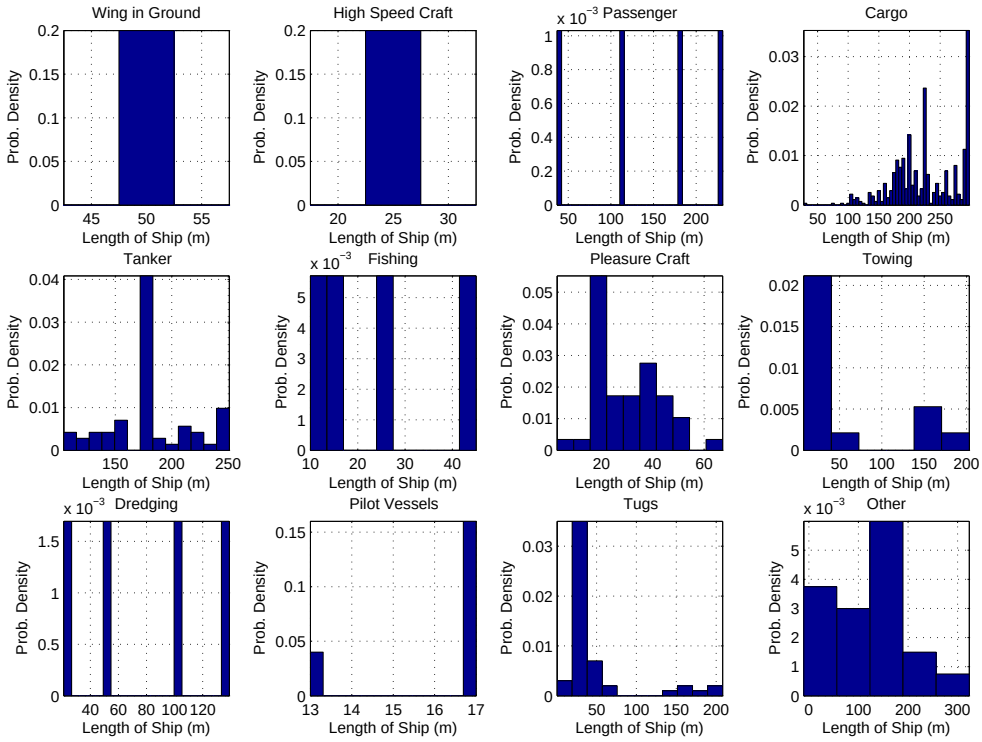

Figure 8: Distribution of reported length of ship by category. Note that the distributions for 'Wing in Ground', 'High Speed Craft', 'Passenger', 'Fishing' and 'Dredging' are unstable due to the limited number of ship in those categories; the distribution in 'Pilot vessels' comes from a number of vessels, but they are all built to mostly the same specification and therefore tend to cluster.

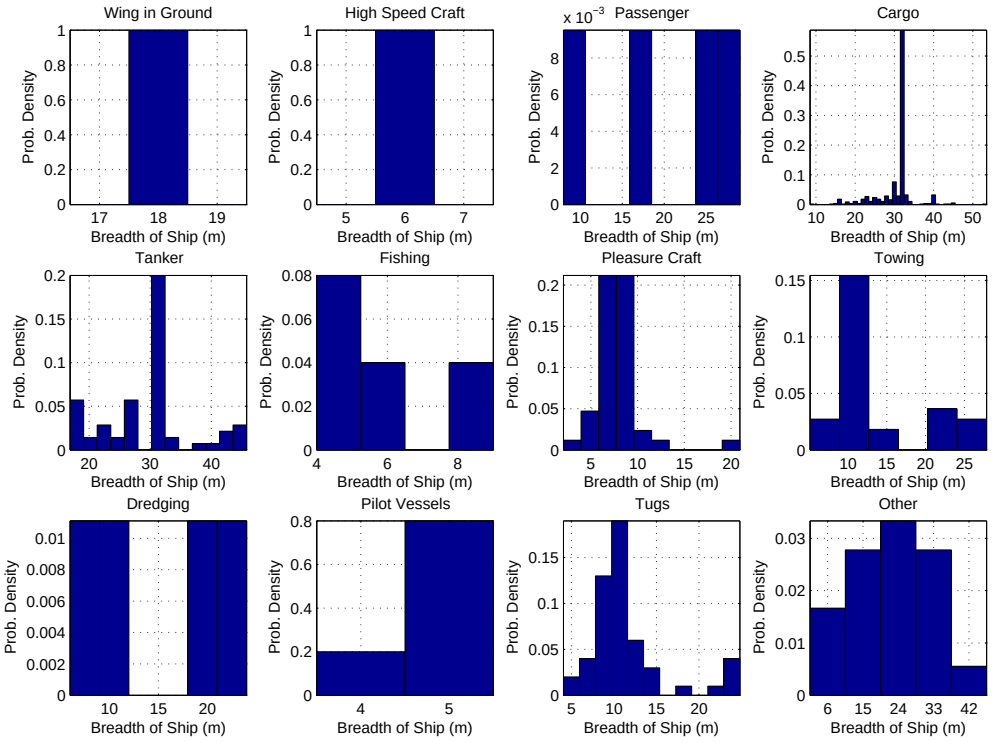

Figure 9: Distribution of reported breadth of ship by category. Refer to Figure 8 for comments on stability of estimation. Note here the prevalence of distinct preferred beam widths due to physical transit limitations in some classes.

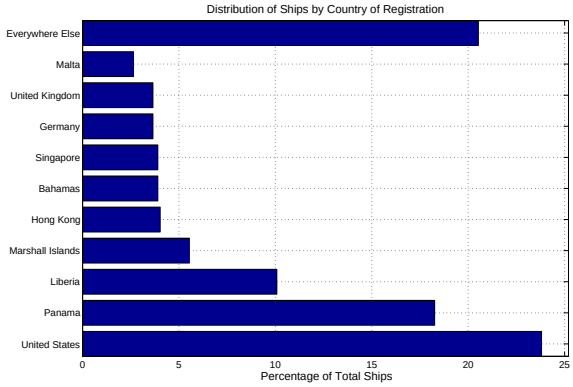

Figure 5: Distribution of ships by country of registration. A total of 42 countries were observed in Norfolk, VA from 2008-10-01 to 2008-11-30, of which the top ten countries are shown. Note particularly the prevalence of 'flags of convenience' in the traffic.

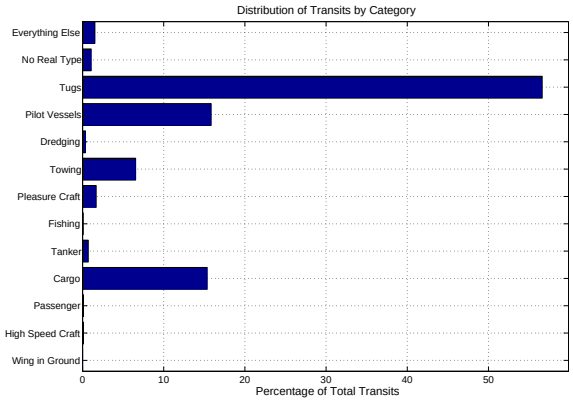

Figure 6: Distribution of detected transits by ship category. A total of 6362 transits were detected in Norfolk, VA from 2008-10-01 to 2008-11-30. The disparity between these results and those of Figure 4 indicates that characterization by ship class or transit density alone is insufficient.

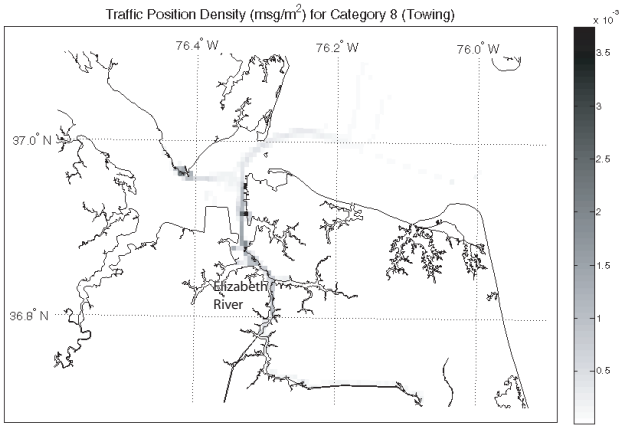

Figure 7: Density of position reports ($\mathrm{msg\,m^{-2}}$) for ships in class 'Towing' (typically tugs). Note the significantly higher density in the Elizabeth River area, and very low density outwith the harbor and main fairways.

ally build a distribution reliably. In practice, we would simulate these by sampling with replacement from the given population (a form of Bootstrap Sampling [18]), although we would prefer to observe over longer periods and/or more areas in an attempt to improve the distribution. There are difficulties in either case: in the former case theoretical, in the latter case physical (e.g., are the distributions the same over larger areas?); this is a matter of on-going research.

Second, we observe somewhat in the length distribution, but much more distinctly in the breadth distribution, a very marked preference for particular dimensions of ships, most particularly in the 'Cargo' and 'Tanker' categories. These are readily explained by the requirement to satisfy PANAMAX or SUEZMAX restrictions on length and breadth ($294\,\mathrm{m} \times 32\,\mathrm{m}$ approximately for PANAMAX and breadth 40 m for SUEZMAX); similar constraints are seen in the draught estimates. This effect is even more distinctive when we consider the distribution at a particular terminal point, Figure 10, a container terminal marked '5' in Figure 11. Here, the dominance of 32 m wide and 294 m long ships indicates that the significant majority are PANAMAX carriers (draught of $< 12\,\mathrm{m}$ is also required). This distribution means that we are likely to have to model PANAMAX carriers as a separate class for those terminal zones that include them, splitting the traffic into two classes: PANAMAX, and 'everything else' according to the remaining distribution. In many cases, the stability of the 'everything else' density is compromised at a particular terminal zone due to the dominance of one particular type of ship. In this case, we can extract

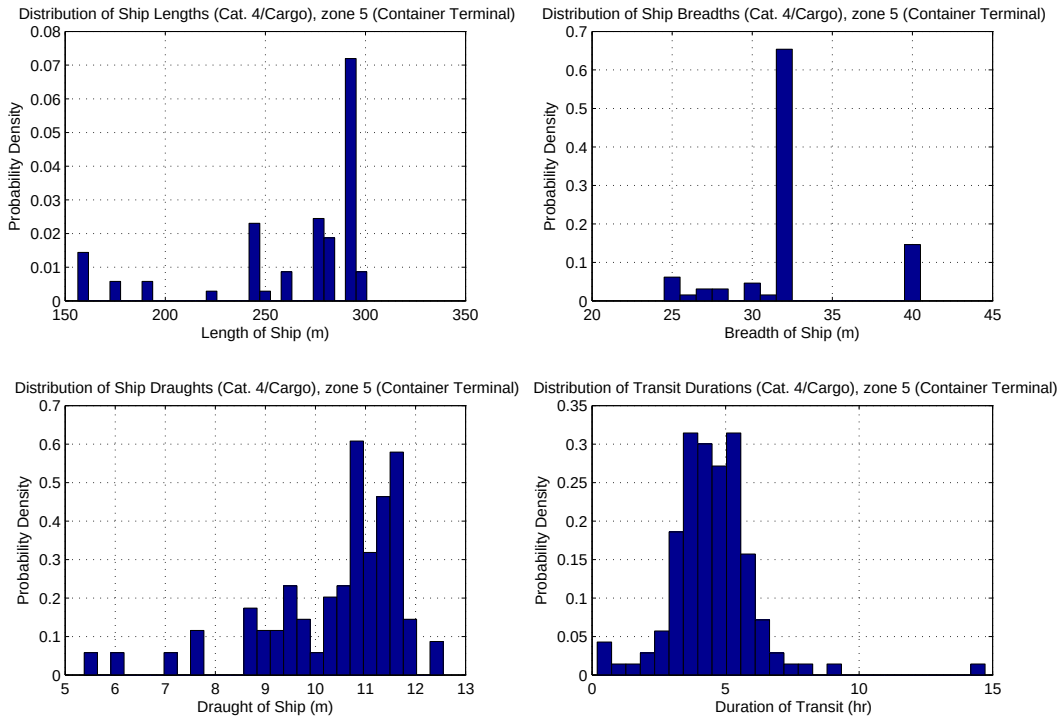

Figure 10: Physical dimensions of cargo ships berthing at zone 4.5 ($76°19'\,\mathrm{W}, 36°55'\,\mathrm{N}$), with transit durations from outside the analysis area. (See Figure 11 for spatial context.)

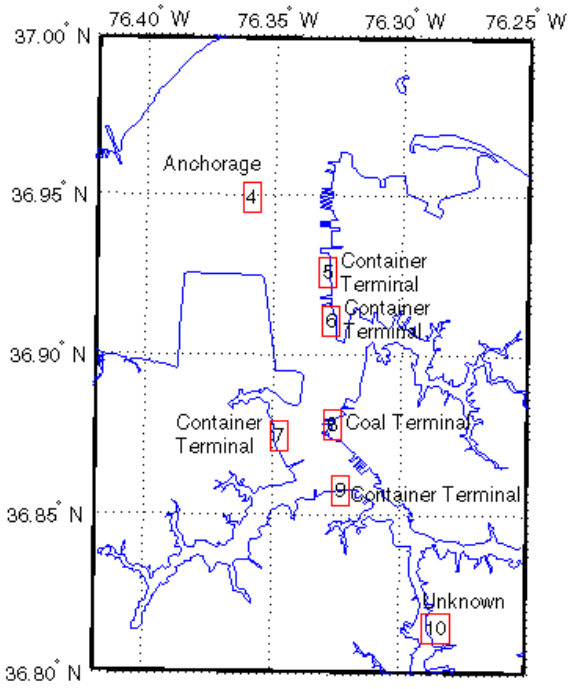

Figure 11: Layout of a subset of the automatically identified terminal zones for cargo ships in Hampton Roads/Norfolk, VA. Figure 10 applies to the container terminal marked '5' in the center of the figure.

those ships in the dominant class, and then agglomerate the remaining ships in the category, e.g., all cargo ships except the PANAMAX carriers, in order to form a more stable estimate.

Finally, we observe that both 'Towing' and 'Tug' classes are bimodal in their lengths and breadths. The cause appears to be that a minority of operators are resetting these dimensions to reflect the size of the vessel that they have in tow. Unfortunately there appears to be some confusion as to when this should occur: some operators are switching from 'Tug' to 'Towing' to indicate their current activity, while some remain 'Tug's, but change their dimensions to suit the tow. For the simulation scheme under development here this is readily resolved: tugs and the ships they service are typically very different sizes. For establishing a consistent navigational and situational awareness in the operations area, however, this observation has significant implications.

3.4 Transit Durations

While underway, the duration of the transit is a large factor in the overall risk that a ship takes in making a voyage: the longer the ship is underway, the more likely it is that something will happen. Understanding the duration of transits is therefore essential in building statistical models of risk.

The overall duration of transits by category of ship is given in Figure 12, where we show the empirical distribution and approximating analytic distributions. For categories where estimation is stable, we observe a number of different regimes. Cargo ships apparently do not delay in getting to their destination as might be expected, so that their transit times are heavily peaked around the mean transit time of 4.3 hr. Tankers, on the other hand appear to have a skewed distribution such that there are a much higher proportion taking longer to come in, possibly due to concerns of safety during docking. Both 'Towing' and 'Tug' categories show distinctly bimodal distributions, corresponding to in-harbor (shorter) and trans-harbor tows. The modes of the trans-harbor tows correspond to the transit times for the cargo ships they are presumably assisting, raising our confidence that this is a real effect.

It is more difficult to explain the behavior of the 'Pleasure Craft' category, which appears to approximately follow an exponential distribution. We hypothesize that this simply reflects the nature of these craft. Most of them are around 8 m long and spend most of their time in the higher reaches of the Elizabeth River, Figure 13, rather than going out to sea. We should therefore expect most trips to be short, with diminishing numbers of longer transits up to a little longer than the nominal transit time for the bigger ships since those craft that do leave the shelter of the harbor have further to go. In fact, we observe many ocean-going super-yachts in the dataset, which may account for the smaller number of longer duration transits.

The decision of whether to model the traffic in a category by identified terminal zones or in aggregate depends strongly on the number of members of the category and the intensity of transits. Lower member counts, or categories that are either not regulated or do not require significant support facilities (e.g., do not need docks with cranes) will result in unstable estimates of terminal zones since they do not tend to congregate. Stratification of the data within these complex zones then results in very diffuse estimates of transit or physical properties. It is arguable, however, that such traffic will not admit an appropriately simple statistical model, and we therefore should accept a little modeling uncertainty in a lower fidelity model in return for a simpler aggregate description that keeps the essence of the category. In this context, we model in detail for 'Cargo', 'Towing', 'Tugs' and 'Tankers', and model in aggregate for all other classes. The aggregate

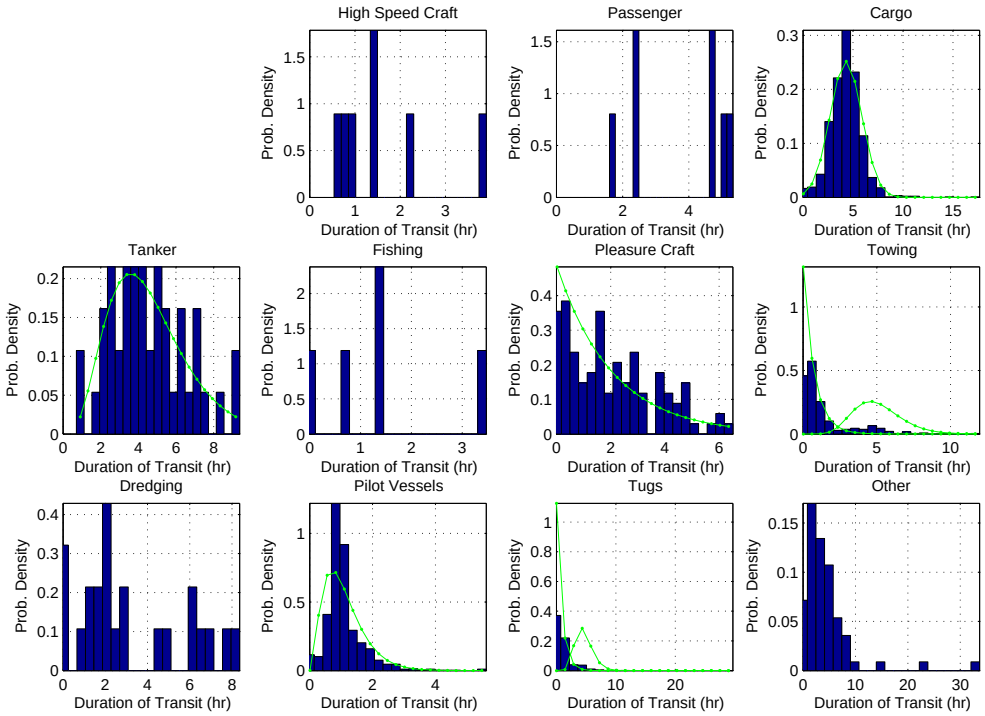

Figure 12: Distribution of transit durations by ship category, with fitted analytical distributions. Estimation is unstable in the 'High Speed Craft', 'Passenger', 'Fishing', 'Dredging' and 'Other' categories, so no models are attempted there. Note that the pairs of distributions fitted for 'Tugs' and 'Towing' categories are individually scaled, and therefore do not reflect the empirical densities, which are composite scaled.

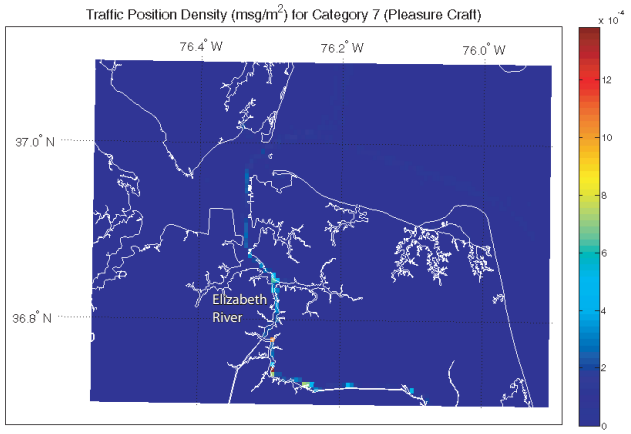

Figure 13: Density of position reports ($\mathrm{msg\,m^{-2}}$) for ships in class 'Pleasure Craft'. Note the prevalence of the counts in the higher reaches of the Elizabeth River area, accounting for the higher proportion of short transits.

classes are represented by a density distribution such as Figure 13, a composite physical dimensions distribution set such as Figure 8 and 9 and a duration of transit distribution such as Figure 12.

3.5 Transit Epochs and Densities

A final question in assessing a simulation model for the larger commercial traffic in the area of interest, where we model in detail, is to assess how often and when the traffic occurs. For each terminal zone identified by the clustering algorithm, we determine those transits that end in the zone separately from those that start there, how many transits are observed in any day, when they arrive and depart, and how long they stay in the zone (the sojourn time). Illustrative results for zone 4.5 (c.f. Figure 10) are shown in Figure 14. We observe that in the container terminal there are 2.13 transits/day on average (i.e., on average one ship per day), which is consistent with the majority of the sojourn times being under 24 hr, and confirmed by the observation that the terminal facility, Figure 15, can likely only service one ship of the sizes indicated in Figure 10 at a time.

More interestingly, the arrival and departure time show some evidence of clustering. The local timezone during this experiment is UTC-5 (Eastern Standard Time), so the early cluster of arrivals are late evening, the second cluster are early morning (local dawn is approximately 1000-1100 UTC during the period of observation) through the early afternoon, and the last cluster correspond to a late afternoon arrival (local dusk is approximately 2200-2300 UTC here). The

Figure 15: Aerial photograph of the zone 4.5 (source: Google Maps/Digital Globe) showing the container terminal characterized in Figure 10 and 14.

extended period of the day arrivals in the middle cluster might be explained by ships that wait at the pilotage area off Cape Henry until first light, and then start their transit. Explanation of the departure times is not as simple, although the same general pattern is observed as for arrival times. This pattern seems to suggest that tidal effects are not strongly observed in the arrival or departure times (a more diffuse pattern without clustering would be expected otherwise), although this is still the subject of ongoing research.

As with all of the other statistics derived here, these behaviors are strongly dependent on the category of traffic and the particular terminal zone. Figure 16, for example, shows the behavior of the pilot boats stationed at Lynnhaven, just west of Cape Henry (Zone 1 in Figure 17), who service the pilotage zone at the entrance to Chesapeake Bay with four boats. The data here show that the pilot boats run essentially all day, and on average 15.9 transits/day, indicating that a very different, but simpler, model would be required to simulate this traffic.

3.6 Density of Class B Transceivers

On 2008-09-19, the Federal Communications Commission (FCC) announced that is would allow Class B devices in the United States and on 2008-10-08, the FCC approved the first US Class B device. Other countries (e.g., Canada) had already been using Class B transceivers by these dates.

Class B transceivers present an opportunity to sam-

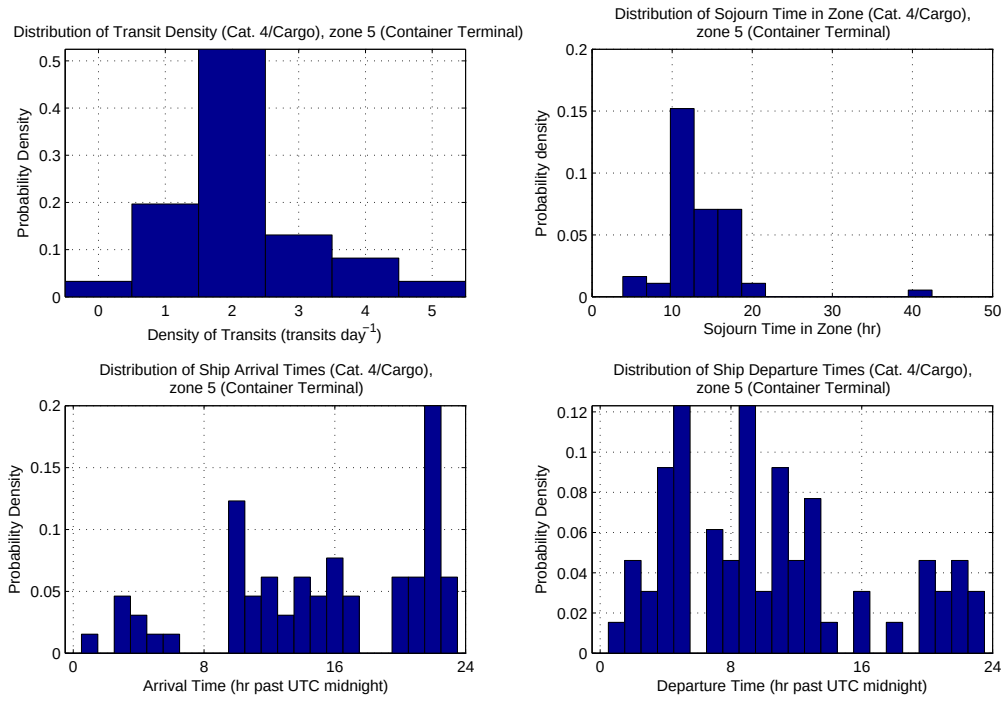

Figure 14: Transit density, arrival and departure times for cargo ships berthing at zone 4.5 ($76°19'\,\mathrm{W}, 36°55'\,\mathrm{N}$). (See Figure 11 for spatial context, and Figure 15 for an overview.) Note that the local timezone is Eastern Standard Time for these data, which is UTC-5.

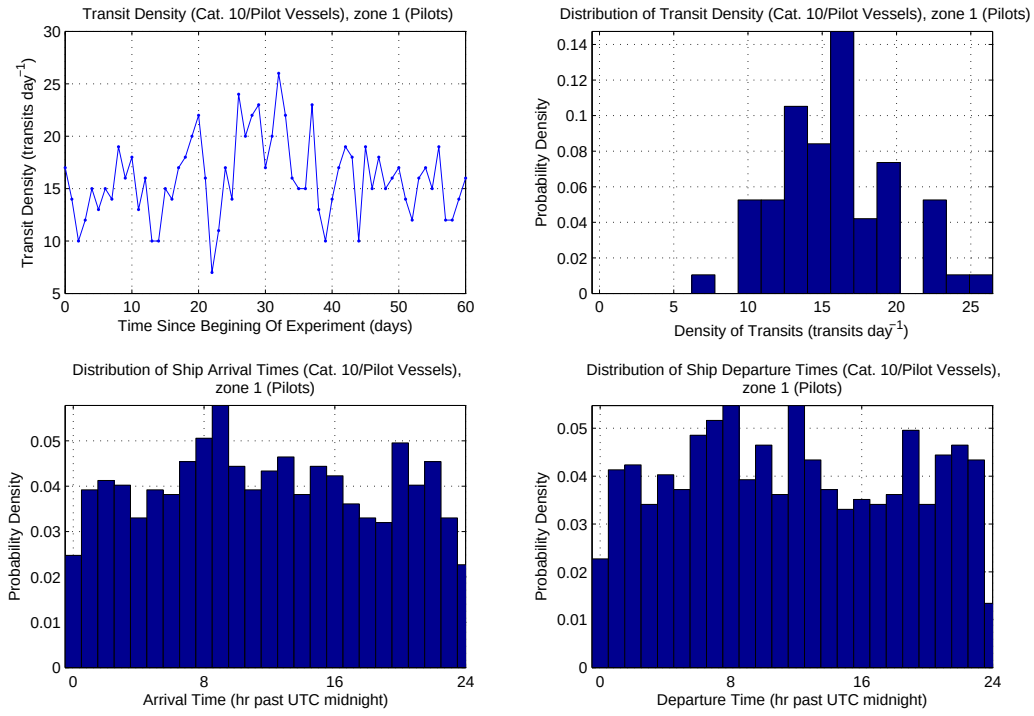

Figure 16: Transit density, arrival and departure times for the pilot boats stationed at Lynnhaven Bay servicing the pilotage area at the entrance to Chesapeake Bay. Note the difference in distribution with respect to Figure 14: pilot boats run all the time because they service traffic from all sources and to all destinations.

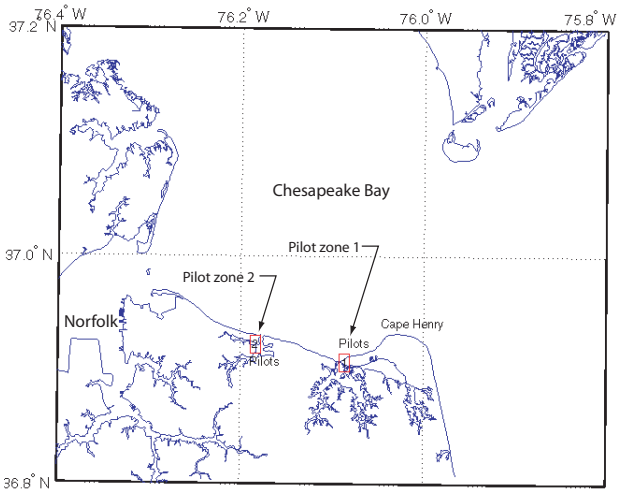

Figure 17: Layout of automatically identified terminal zones for pilot boats in the Chesapeake Bay area. Figure 16 corresponds to zone 1, just west of Cape Henry.

ple the traffic of very different types of vessels compared to Class A, that are likely to cover different regions of navigable waters, especially areas outside of established channels. We examined 459 days from 2008-01-01 to 2009-04-04 for both Class A and Class B vessel reports from basestation b003669708 in the San Francisco, CA area. For this test, we used the CCOM/JHC N-AIS 1 message/vessel/minute research and development feed. Being a development feed, it suffered from a number of small outages that can be seen as drops in the Class A position reports (Figure 18).

Table 4 compares Class B position reports from unique vessel identities per day against Class A. The Class A position messages demonstrate that vessel traffic in the harbor is relatively constant, with a traffic mixture of US and foreign MMSIs averaging around 28,830 position reports per day. Both Class B position reports and unique MMSIs are more than two orders of magnitude lower than Class A.

The table of means per day hides an important trend in the data. Figure 18 shows Class B position reports exhibit a switch from foreign registered MMSIs to US registered MMSIs after the FCC approval. Figure 19 better illustrates the trend in the number of Class B transceivers installed in the San Francisco area. On the peak day, eight Class B transceivers were seen. This trend demonstrates the growth of the Class B population. While not yet significant, there is a strong upward trend as mariners adopt this technology that has just recently become available. However, without more Class B position reports, it is difficult to make any kind of analysis of vessel movements.

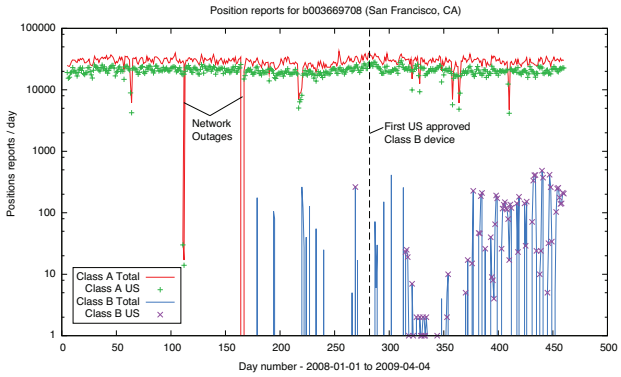

Figure 18: N-AIS 1 position/minute/vessel position reports delivered to CCOM/JHC from a basestation in the San Francisco, CA area. The large swings in the number of position reports are likely loss of messages due to the architecture of the data logging system. Class A position report traffic is roughly constant with approximately 2/3 of the reports being from US Vessels. US Class B transceivers appear after Class B devices were approved in October, 2008. The Class B traffic is becoming more regular and gradually increasing.

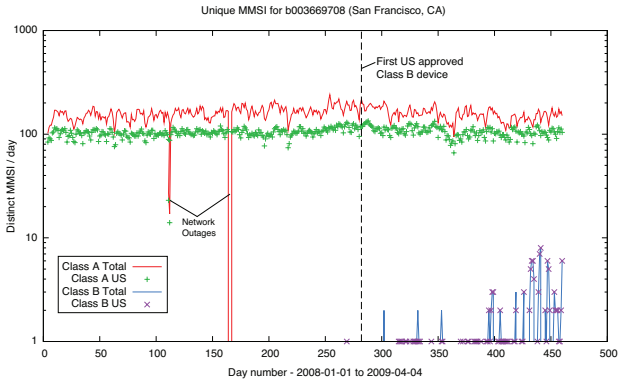

Figure 19: Unique ship identities seen based on MMSI reported in N-AIS 1 position/minute/vessel position reports delivered to CCOM/JHC from a basestation in the San Francisco, CA area. There are non-US transceiver active before the FCC approved Class B transceivers.

Class/Mean	P_{tot}	P_{us}	$MMSI_{tot}$	$MMSI_{us}$	Total MMSI
Class A	28830	20170	159	104	**3366**
Class B	100	80	1.7	1.4	**25**

Table 4: Mean for days with Class N traffic of position reports and unique MMSI/day for station b003669708 in San Francisco, CA. *Total MMSI* is the number of the unique MMSI seen across the entire timespan.

4 Discussion

4.1 Retrospective AIS Traffic Analysis

AIS has been the subject of a number of papers that have attempted to harvest data for retrospective analyses (e.g., [2, 10, 20, 21] among many others). It is not, however, ideal for these sorts of analyses. To be fair, the system was not designed for it, so this is perhaps understandable. The utility of the system for this task could be greatly improved, however, with some simple modifications.

First, the protocol lacks an obvious sequence number similar to the design of Transmission Control Protocol (TCP) [22]. Although reliable end-to-end communication is clearly never going to be a design goal of a broadcast system such as AIS, having a sequence number that simply incremented after each packet was broadcast would allow duplicate packets to be more reliably detected, making analysis and tracking significantly simpler. The number of bits to use for this depends on the available packet space and transmission rate. Since data packets are typically sent out at low rates [3, Annex 1, §4.2.1], however, sequence numbers as low as four bits would probably be sufficient. (It may also be possible under some circumstances to use the synchronization state information for the same goal, which we are currently investigating.)

Second, the system does not assign a source timestamp in every packet indicating the time of validity for the data. Most Class A transceivers have a very good sense of time in the GPS receiver attached to them, and use this to synchronize themselves into the SOTDMA scheme that AIS implements [3, Annex 2, §3.3.4.4], however, and providing this information would result in reliable sequencing of information and subsequent analysis. The information can be gleaned at several minute intervals from the synchronization state information in each packet, but it is difficult to interpolate the information between these events in the face of packet loss. Time within the day would entail 11 bit packet additions; time within the hour 6 bit additions. One could argue that this information could be inferred on shore when the data is recorded. However, delays in the medium access algorithm at the transceiver, during reception and in transfer to the logging computer make this difficult even under ideal conditions. Timestamping data at source always works better.

Third, we observed that the data stream has an unexpected number of packets with what appear to be undetected random bit errors. The system uses CRC-16 [23] to compute an error detecting code, and therefore should identify any burst errors up to approximately 10% of the payload data length for AIS. It is uncertain whether the observed packet corruption is a product of lax implementation of this scheme at a particular receiver or evidence of significantly higher bit error rate than this, but some further investigation is probably indicated to avoid these packets getting to the Application Layer of the protocol stack.

Finally, many of the issues that we had to resolve before performing detailed analysis of the data arise because of misconfiguration either of the static data of the ship (name, IMO or MMSI number, etc.) or inappropriate use of the dynamic data (ship type code, dimensions or 'Navigation Status' flag). This is almost surely evidence of some confusion on the part of the user community of how to interpret the meanings of some of the parameters (e.g., 'Do dimensions mean just me, or me and the vessel I have in tow?', or 'Do I use *fishing* to mean that I am a fishing vessel, or that I am currently fishing?'), but some of this probably reflects an inherent characteristic of AIS. It seems evident that there is still some work to be done in training of users, and in clarification of intent through guidance from the appropriate authorities, before the AIS data stream becomes reliable for this sort of analysis without a great deal of data checking and filtering.

4.2 Characterization of Traffic

Our preliminary analysis of AIS data for a 60 day period appears to show that there is sufficient information for a number of different analyses of the traffic in a given area, at increasing levels of complexity. The difficulty that we face, however, is to identify the appropriate level at which to conduct the analysis for each class of shipping or activity.

In the case of large commercial traffic and their attendant services (e.g., tugs and pilots), Figures 10, 12 and 14 show that we can construct a fairly detailed

model of the traffic behavior in the area.

In the case of smaller vessels and those with less organized behaviors in general, it is unlikely that we will ever achieve this level of detail, irrespective of the amount of data that is available. In these circumstances, we expect that the model will have to be formulated using an estimate of the range of the vessels (e.g., Figure 13) and their general characteristics, Figures 8, 9 and 12.

The extent to which this lack of detail will influence the uncertainty analysis that is our ultimate goal remains to be seen. On the one hand, smaller vessels are typically less constrained by their dimensions, and therefore have a less critical time moving in most areas. In addition, when they do have an incident, the relative costs (in the most general sense) tend to be somewhat less than for large commercial traffic, and therefore the loss of detail may not be too significant in assessments intended to apply to the whole area.

On the other hand, however, where small craft interact with larger craft, or each other, the lack of detail might be significant. For example, the effects of smaller ships on the behaviors of larger vessels could impact the assessment of the uncertainty associated with the main users of the area. How significant this effect is likely to be is a matter of current research.

4.3 Class B Transceivers

While the Class B dataset presented here is too sparse for useful analysis, Class B transceiver reports hold promise for adding vessels with very different coverage of areas that we would like to evaluate. Additionally, the requirement that the vendor of Class B units program the devices is a step towards providing more reliable ship size and draught parameters. However, the mariner must still correctly communicate the vessel parameters and there is still a probability for data entry errors, which are now harder to correct. The 2 W transmit power of these transceivers and the expected lower position of the antenna may drastically reduce the ability to successfully receive the position reports. Finally, the reduced reporting rate of position messages will introduce more positioning errors for vessels that are maneuvering.

4.4 Simulation and Risk Analysis

The idea of risk models to assess such things as harbor channel availability [24, 25] and resurvey potential [2] have been proposed previously; we have previous applied the same principles to the more general problem of finding an uncertainty description for a given area [1]. (Other models such as per-transit analysis, path planning, resurvey potential and survey monitoring are of course possible and enabled by the same methods.)

Much of the difficulty with such models is obtaining appropriate calibration data for such factors as ship types, densities, routes, etc. for a particular area. The current work shows that such parameters are extractable from AIS traffic with sufficiently robust processing. Knowing where ships are heading, and how often, it is relatively simple to construct direct simulations of their traffic patterns. For example, we could model cargo traffic by selecting a destination from the identified zones according to their transit intensity, an arrival time from that zone's distribution and a sojourn time at the dock. We then have sufficient information to run a Monte Carlo [26] analysis of the transit and thereby assess the potential risk to the ship in this process. Repeated sufficiently, we can build up statistics for likely risk in the area, matched to the classes of traffic and their individual behaviors.

As appealing as this idea sounds, however, it suffers from a difficulty found in any simulation system based on empirical data: the system will only predict that which it has already seen. While this is sufficient to analyze risk to the shipping that is occurring today, it does not assist in predicting what might happen in the future—or what might happen when things go wrong (e.g., when a ship has to leave the channel). If we are to assess the uncertainty for the whole area of interest then we need to consider methods to extrapolate the observed behaviors plausibly to the remainder of the area, and to consider the potential for unexpected events. How to do this effectively is still an open, and very interesting, question.

5 Conclusions

We have considered the problem of robustly mining AIS data for information suitable to calibrate risk models for ships in a circumscribed area. We have shown that although AIS data has a number of difficulties for such retrospective analyses, it is possible to extract models of ship behavior in an area at a suitably nuanced level to allow for simulation based models to be partially calibrated. The level of characterization possible depends on the type of traffic, and to some extent on the number of observations that remains after the data is stratified sufficiently to elicit the underlying behaviors. In some cases, we expect that less detailed models will be generated, constrained by the very complex dynamics of the populations of users, and/or the sparsity of data.

We aim to assess risk to the mariner associated with operating in an area, given the available sources of data and their constituent uncertainties. The current work informs this effort, but it is an open question as to how we might extrapolate the observed data to allow for the unexpected or rare events that characterize high-risk situations on the water. We hope these questions will be answered in future research.

Acknowledgements

The support of NOAA grant NA05NOS4001153 for this work is gratefully acknowledged by the authors. We would also like to thank the USCG RDC for their extensive technical discussion and support which contributed to the success of this study.

References

[1] B. R. Calder. Uncertainty representation in hydrographic surveys and products. In *Proc. 5th Int. Conf. on High Resolution Survey in Shallow Water, Portsmouth, NH*, Oct 2008.

[2] M. Lundkvist, L. Jakobsson, and R. Modigh. Automatic Identification System (AIS) and Risk-Based Planning of Hydrographic Surveys in Swedish waters. In *Proc. FIG Working Week 2008*, June 2008.

[3] International Telecommunications Union. *Technical characteristics for an Automatic Identification System using Time Division Multiple access in the VHF maritime mobile band*, 2007. (ITU Recommendation: ITU-R M.1371-3).

[4] U.S. Coast Guard Navigation Center N-AIS Data Feed and Data Request. (Online, http://www.navcen.uscg.gov/enav/ais/disclaimer.htm).

[5] International Association of Marine Aids to Navigation and Lighthouse Authorities. *The Automatic Indentification System (AIS): Operational Issues*, 2004. (IALA Guideline No. 1028, Ed. 1.3).

[6] International Association of Marine Aids to Navigation and Lighthouse Authorities. *The Universal Automatic Identification System (AIS): Technical Issues*, 2002. (IALA Guidelines, Edition 1.1).

[7] International Association of Marine Aids to Navigation and Lighthouse Authorities. *The Management and Monitoring of AIS Information*, 2005. (IALA Guideline No. 1050).

[8] International Association of Marine Aids to Navigation and Lighthouse Authorities. *Automatic Identification System (AIS) Shore Station and Networking Aspect relating to the AIS Service*, 2005. (IALA Recommendation A-124).

[9] A. Marati-Mokhtari, A. Wall, P. Brooks, and J. Wang. Automatic Identifcation System (AIS): Data reliability and human error implications. *The Journal of Navigation*, 60:373–389, 2007.

[10] K. Naus, A. Makar, and J. Apanowicz. Using AIS data for analyzing ship's motion intensity. In *Proc. 7th International Symposium on Navigation (Gydnia, Poland)*, 2007.

[11] K. B. Williams. *RFP for the Nationwide Automatic Identification System Increment 2, Phase 1.* U.S. Coast Guard, 2007. (Rev. 1.0).

[12] D. R. Hipp. *The SQLite C/C++ API, version 3.* Hwaci Aplied Software Research, 2008. (Online, http://www.sqlite.org/c3ref/intro.html).

[13] Refractions Research. *PostGIS User Manual*, 2008. (Online, http://postgis.refractions.net/documentation/manual-1.3).

[14] National Marine Electronics Association. *Standard for Interfacing Marine Electronic Devices*, 1992. (NMEA Standard 0183, Rev. 2.0).

[15] L. Luft and J. Spaulding. Specification for the National AIS system Timestamps and Metadata - format 0. Technical report, U.S.C.G Research and Development Center, 20 April 2006.

[16] J. Seward. *BZIP2: A program and library for data compression*, 2008. (Online, http://bzip.org/1.0.5/bzip2-manual-1.0.5.html).

[17] K. Schwehr. The `noaadata-py` Software Toolset, v0.42, 2009. (Online, http://vislab-ccom.unh.edu/ schwehr/software/noaadata).

[18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95284-5.

[19] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2009. (Online, http://www.scipy.org).

[20] L. Hatch, C. Clark, R. Merrick, S. Van Parijs, D. Ponirakis, K. Schwehr, M. Thompson, and

D. Wiley. Characterizing the relative contributions of large vessels to total ocean noise fields: A case study using the Gerry E. Studds Stellwagen Bank National Marine Sanctuary. *Environmental Management*, 42(5):735–752, 2008.

[21] M. Gucma. Combination of processing methods for various simulation data sets. In *Proc. 7th International Symposium on Navigation (Gydnia, Poland)*, 2007.

[22] V. G. Cerf and R. E. Kahn. A protocol for packet network intercommunication. *IEEE Trans. Comm.*, 22(5):637–648, 1974.

[23] ISO/IEC JTC1/SC6. *High-level Data Link Control (HDLC) Procedures: Frame Structure*. International Organization for Standardization, 1993. (ISO/IEC Standard 3309: 1993).

[24] A. L. Silver and J. F. Dalzell. Risk-based decisions for entrance channel operation and design. *Int. J. Offshore and Polar Eng.*, 8(3):200–206, 1998.

[25] L. Gucma and M. Schoeneich. Probabilistic model of underkeel clearance in decision making process of port captain. In *Proc. 7th International Symposium on Navigation (Gydnia, Poland)*, 2007.

[26] C. P. Robert and G Casella. *Monte Carlo Statistical Methods*. Springer texts in Statistics. Springer-Verlag, New York, 2004. ISBN 0-387-21239-6.